PREPARING FOR INFLUENZA SEASON
INTERIM REPORT

Project Overview

- **Motivation:** The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.

- **Objective:** Determine when to send staff, and how many, to each state.

- **Scope:** The agency covers all hospitals in each of the 50 states of the United States, and the project will plan for the upcoming influenza season.

Hypothesis

- Elderly have a higher chance of dying from Influenza. (If 65+ years old then higher chance of dying from Influenza)

Data Overview

- Population data by geography US Census data
    - This data set contains the total US population per state including segregations by male, female and age groups ranging from year 2009 till 2017
- Influenza deaths
    - The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017. Counts are broken into two categories: state and age.

Data limitations

- Population data
    - Limitation is that the data is not collected only once a year. Certain lag in data up to date status.
    - Time range of the dataset could be more recent. Only reaches till 2017.
    - Population numbers are estimates.
- Influenza deaths
    - Death counts of age groups 1 till 74 are listed as “Suppressed”. Affecting over 80% of the death data.

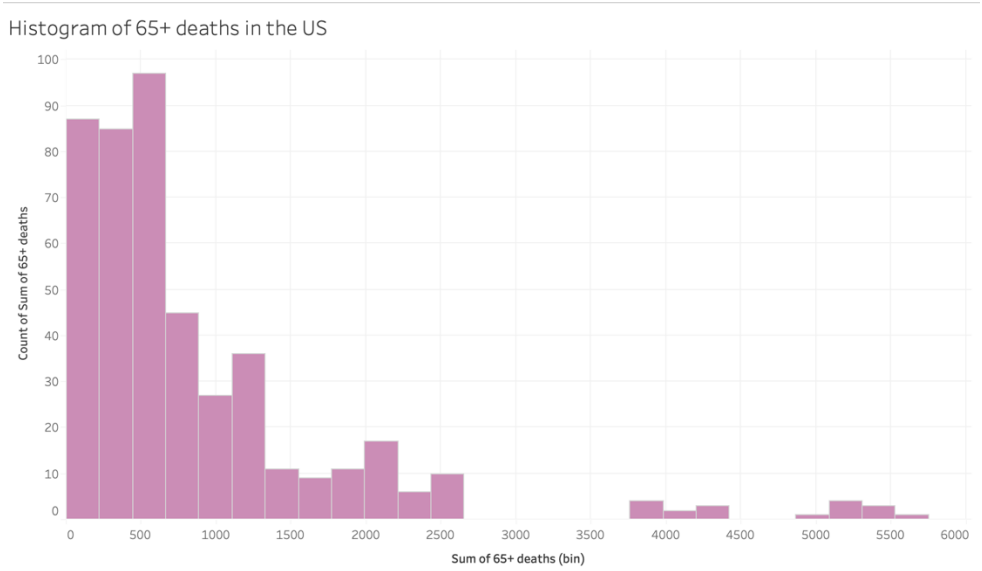
Descriptive Analysis

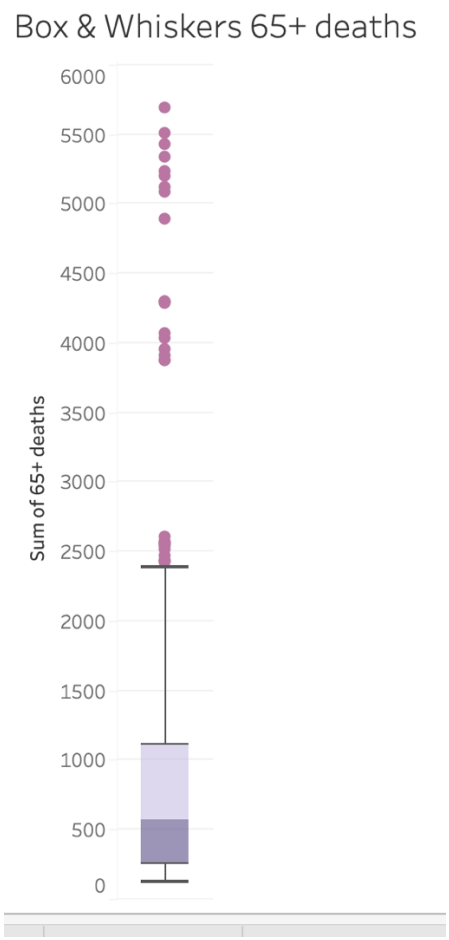
In this section I have been diving deeper in the descriptive analysis. The main reason herefore is to test my hypothesis on its significance and confirmation of its valuation. In the table below you can find some variables which are the basis to start testing the hypothesis. To make it more clear I added some visualization below too.

- Overview of core variables

Variable name	65+ Influenza deaths	65+ Population
Variance	952508	786893953996
Deviation	976	887070
Sample or population	Sample	Sample
Mean	890	807370
Outlier SD 1	1866	1694440
Outlier SD -1	-86	-79700
Outlier SD 2	2842	2581511
Outlier SD -2	-1062	-966771
Outlier SD 3	3818	3468581
Outlier SD -3	-2038	-1853841
Outliers	18	12
Count	459	459
Outliers in %	3.92%	2.61%

Visualization of 65+ Influenza deaths variables with help of a Histogram and box and whisker chart:

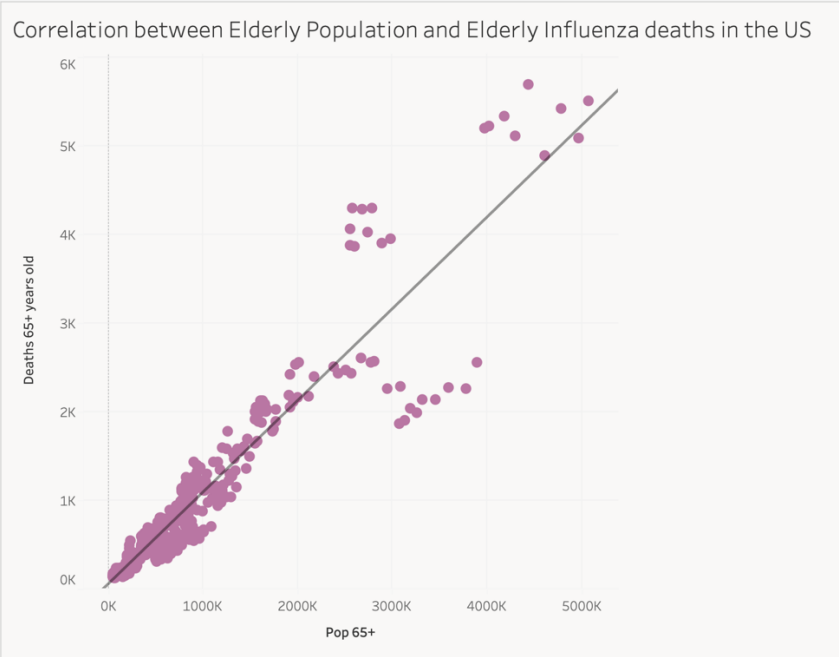




Correlations

Variables	**Influenza deaths 65+ / Total population 65+**
Correlationship explained	To test if there is a relation between the deaths of elderly and the elderly population.
Correlation coefficient	0.940791018
Strong / weak correlation	Very strong
Useful? Does it help the hypothesis	This was actually expected but now proven that there is a very strong relationship between elderly and death rates in the elderly population. This will be very useful for proving the hypothesis, especailly to prove that certain states with higher elderly population will be more exposed to the influenza deaths.

To understand the table above I will show a visualization of the relationship between the two variables “Influenza deaths of 65+” and “Population of 65+”.



You can see above that the data is very closely scattered around the trend line which refers to a very strong correlation of these two variables.

Which gives us the conclusion of the insight mentioned here below.

Results and Insights

Hypothesis	Elderly have a higher chance of dying from Influenza If 65+ years old then higher chance of dying from Influenza
Null Hypothesis	the death rate of 65+ years of age is equal or less than people of 64 years or less
Alternative Hypothesis	the death rate of 65+ years of age is higher than people of 64 years or less
one-tailed or two-tailed test	Since we are only interested in the deaths of the elderly its a one tailed test. We will test if the change of the independent variable will influence the dependent variable.
significance level, Alpha	0.05

After implementing a t-test the outcome of a p-value of 4.886R-195, which is the closest you can get to 0 and smaller than the significance level Alpha of 0.05 we can conclude that we can reject our null hypothesis. In other words we can say with a confidence of 100% that elderly have a much higher chance of dying from Influenza than other age groups.

Next steps

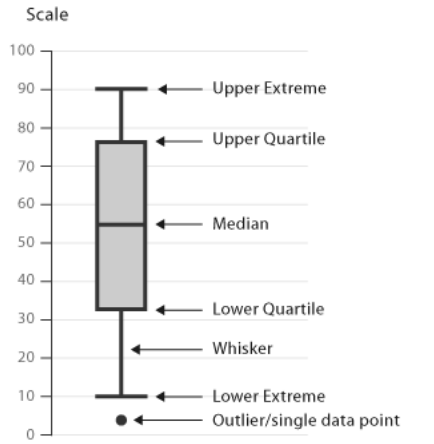
- Use tested hypothesis to implement a heat map for distributing medical staff accordingly.
- Prepare spatial and temporal visualizations
- Prepare composition & comparison charts
- Prepare statistical visualizations
- Storytelling with data presentations
- Present findings to stakeholders in final video presentation

Appendix

- Influenza death data set source: CDC
- Population data set source: US Consensus Bureau

Explanations:

- Hypothesis:
    - A hypothesis is an assumption, an idea that is proposed for the sake of argument so that it can be tested to see if it might be true.
- Descriptive Analysis:
    - a sort of data research that aids in describing, demonstrating, or helpfully summarizing data points so those patterns may develop that satisfy all of the conditions of the data. It is the technique of identifying patterns and links by utilizing recent and historical data.
- Variance:
    - The simple definition of the term variance is the spread between numbers in a data set. Variance is a statistical measurement used to determine how far each number is from the mean and from every other number in the set.
- Deviation:
    - a measure of how dispersed the data is in relation to the mean. Low, or small, standard deviation indicates data are clustered tightly around the mean, and high, or large, standard deviation indicates data are more spread out.
- Sample or population:
    - A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population. In research, a population doesn't always refer to people.
- Mean:
    - The "average" number; found by adding all data points and dividing by the number of data points.

- Outlier:
    - An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
- Box and Whisker plot:
    - 

- Correlations Coefficient:
    - a number between -1 and 1 that tells you the strength and direction of a relationship between variables. In other words, it reflects how similar the measurements of two or more variables are across a dataset.2 Aug 2021
- Null Hypothesis:
    - This can be thought of as the implied hypothesis. “Null” meaning “nothing.” This hypothesis states that there is no difference between groups or no relationship between variables. The null hypothesis is a presumption of status quo or no change.
- Alternative Hypothesis:
    - One that states there is a statistically significant relationship between two variables. It is usually the hypothesis a researcher or experimenter is trying to prove or has already proven.
- One-tailed and two tailed test:
    - A one-tailed test results from an alternative hypothesis which specifies a direction. i.e. when the alternative hypothesis states that the parameter is in fact either bigger or smaller than the value specified in the null hypothesis.
    - The main difference between one-tailed and two-tailed tests is that one-tailed tests will only have one critical region whereas two-tailed tests will have two critical regions.
- Significance level:
    - Alpha is also known as the level of significance. This represents the probability of obtaining your results due to chance. The smaller this value is, the more “unusual” the results, indicating that the sample is from a different population than it's being compared to, for example. Commonly, this value is set to 0.05.